LBNL-59384

The complete chloroplast genome sequence of *Pelargonium* × *hortorum*: Organization and evolution of the largest and most highly rearranged chloroplast genome of land plants

Timothy W. Chumley[1], Jeffrey D. Palmer[2], Jeffrey P. Mower[2], H. Matthew Fourcade[3], Patrick J. Calie[4], Jeffrey L. Boore[3,5] and Robert K. Jansen[1]

[1]The University of Texas at Austin, Austin, TX USA

[2]Indiana University, Bloomington, IN USA

[3]DOE Joint Genome Institute and Lawrence Berkeley National Laboratory, Walnut Creek, CA USA

[4]Eastern Kentucky University, Richmond, KY USA

[5]University of California, Berkeley, CA USA

Abstract

The chloroplast genome of *Pelargonium* × *hortorum* has been completely sequenced. It maps as a circular molecule of 217,942 bp, and is both the largest and most rearranged land plant chloroplast genome yet sequenced. It features two copies of a greatly expanded inverted repeat (IR) of 75,741 bp each, and consequently diminished single copy regions of 59,710 bp and 6,750 bp. It also contains two different associations of repeated elements that contribute about 10% to the overall size and account for the majority of repeats found in the genome. They represent hotspots for rearrangements and gene duplications and include a large number of pseudogenes. We propose simple models that account for the major rearrangements with a minimum of eight IR boundary changes and 12 inversions in addition to a several insertions of duplicated sequence. The major processes at work (duplication, IR expansion, and inversion) have disrupted at least one and possibly two or three transcriptional operons, and the genes involved in these disruptions form the core of the two major repeat associations. Despite the vast increase in size and complexity of the genome, the gene content is similar to that of other angiosperms, with the exceptions of a large number of pseudogenes as part of the repeat associations, the recognition of two open reading frames (*ORF56* and *ORF42*) in the *trnA* intron with similarities to previously identified mitochondrial products (*ACRS* and *pvs-trnA*), the loss of *accD* and *trnT-GGU*, and in particular, the lack of a recognizably functional *rpoA*. One or all of three similar open reading frames may possibly encode the latter, however.

Introduction

The recent explosion of chloroplast genomic data has confirmed what had earlier been demonstrated through many restriction site mapping studies, that gene content, gene order, and genome structure are largely conserved (Palmer, 1991; Raubeson and Jansen, 2005). These observations are particularly extensive regarding the chloroplasts of most terrestrial plants, especially angiosperms, owing to their extensive sampling. The tobacco chloroplast genome (Shinozaki et al., 1986), as the first to be completely sequenced, is most often the model against which newly sequenced angiosperm genomes are compared, and it is indeed typical of the majority of angiosperms in its length, structural partitions, and relative sizes, gene content, and gene order. In most land plants, chloroplast DNA is a molecule of 120-160 thousand base pairs (kb) with an inverted repeat (IR) of 20-28 kb separating two single copy regions of 80-90 kb and 18-27 kb [the so-named large and small single copy regions (LSC and SSC, respectively)]. The genome usually encodes four rRNAs, 30 tRNAs, and about 80 proteins; the IR typically contains the four rRNA genes and 10-15 other genes.

Deviations from the conserved gene arrangement are typically the result of either changes in the extent of the IR or inversions (Palmer, 1991; Raubeson and Jansen, 2005). Many of these are small changes of one to several hundred nucleotides commonly found at the IR boundaries [although they can be much larger, e.g., 12 kb in *Nicotiana acuminata* (Goulding et al., 1996), 11.5 kb in Berberidaceae (Kim and Jansen, 1994), xx kb in *Lobelia thuliana* (Knox and Palmer, 1999)]. Large inversions are occasionally found as in Asteraceae (22.8 kb) (Jansen and Palmer, 1987; Kim et al., 2005), *Oenothera* (54 kb) (Hachtel et al., 1991; Hupfer, 2000), and Fabaceae (50 kb) (Bruneau et al., 1990;

Doyle et al., 1996; Palmer et al., 1988). These rearrangements are usually limited to a single event or a simple series of events (Downie and Palmer, 1992; Palmer, 1991), as in the Ranunculaceae (Hoot and Palmer, 1994; Johansson, 1999; Johansson and Jansen, 1991; Johansson and Jansen, 1993) and Poaceae (Doyle et al., 1992; Howe et al., 1988; Katayama and Ogihara, 1996).

Complex rearrangements involving multiple events are quite rare, but examples have been identified among conifers (Lidholm et al., 1988; Raubeson and Jansen, 1992; Strauss et al., 1988; Wakasugi et al., 1994), legumes (Milligan et al., 1989; Palmer et al., 1988; Palmer and Thompson, 1981), campanuloids (Cosner et al., 1997; Cosner et al., 2004) and the related lobelioids (Knox et al., 1993), and geraniums (Palmer et al., 1987). Of these, the only ones that are known to be completely sequenced are from two pines (Wakasugi et al., 1994)(Noh, et al., 2003, unpublished, accession NC_004677). The sequenced genome of the parasite *Epifagus* can also be considered highly rearranged, but its rearrangements are mostly due to the large deletions that have severely reduced its genome, and otherwise rearranged only by a single small inversion (Wolfe et al., 1992). Each of these groups may have much to teach us about the pattern, mode, and mechanisms of genome evolution in the chloroplast (Palmer, 1990).

In this study, we present the complete nucleotide sequence of the chloroplast genome of the common garden geranium (*Pelargonium* × *hortorum* L. H. Bailey; Geraniaceae) and compare it to other closely related genomes. This genome was previously found to be unusually large and highly rearranged (Palmer et al., 1987). This early study estimated the genome size to be about 217 kb, or about 40% larger than usual, and concluded that most of this size increase was the result of a three-fold increase in the size of the IR, with

consequent reduction of both single copy regions. Gene order was found to be highly rearranged relative to tobacco; a minimum of six inversions were hypothesized in addition to the aforementioned tripling of the IR size. Two families of dispersed repeats [later characterized as potentially novel DNA (Palmer, 1991)] were detected. These novelties also appear to have contributed to the genome expansion, and recombination between them was proposed as a possible (likely???) cause of the inversions.

Materials and Methods

Methods for DNA isolation, sequencing, and analysis have been described previously (Jansen et al., 2005), but a brief summary is provided here. Commercially available plants of *Pelargonium* × *hortorum* cv. 'Ringo White' (Mower s.n., 4 Sept. 2003 (TEX)) were obtained locally and grown in a greenhouse. Purified chloroplast DNA was isolated with a modified DNAse I method (Kolodner and Tewari, 1972) from 500 g of fresh leaf tissue taken from several plants. The isolated DNA was sheared by repeated passage through a narrow aperture using a Hydoshear device (Gene Machines), then these fragments were end-repaired, gel isolated, and ligated into pUC18 to create a DNA library. These clones were introduced into *E. coli* by electroporation and plated onto nutrient media with antibiotic selection. Resulting colonies were randomly selected and processed robotically for sequencing from each end of each clone using Big Dye (Applied Biosystems) chemistry on an ABI 3730 XL. Detailed protocols are available at <http://www.jgi.doe.gov/sequencing/protocols/protsproduction.html>. A total of 4,608 sequencing reads were generated, which were processed with PHRED and assembled with PHRAP (Ewing and Green, 1998; Ewing et al., 1998). The quality of sequencing

reads and the assembly were verified by eye with Consed 13.0 (Gordon et al., 1998) and Sequencher 4.2 (Gene Codes Corp., 2003).

Gaps that remained in the assembled draft sequence were filled by primer walking on either cloned or PCR amplified templates. No sequences from the SSC region were present in the draft sequence and so it was necessary to use a PCR strategy to sequence through this entire region. Because of difficulties in assembling repeated regions from random reads, each of the IR boundaries were verified by sequencing across them. In all, approximately 20 kb of additional sequencing was necessary to complete the genome. All primer sequences are shown in Table 2, supplemental data.

Upon completion of sequencing and final assembly, genes were annotated using DOGMA (Wyman et al., 2004), supplemented by using direct BLAST comparisons (Altschul et al., 1990). Annotations are based on nucleotide and amino acid similarity and are not experimentally verified. Additional open reading frames were assessed using EditSeq 5.06 (DNASTAR Inc., 2003) and NCBI's OrfFinder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html). Initial ORF searches were limited to frames of 99 bp or longer, and only those with BLAST hits to genes of known function or recognized ORFs were considered further.

Exact microsatellite repeats were examined using Msatfinder ver. 1.6.8 (Thurston and Field, 2005) with thresholds of seven repeat units for mononucleotide simple sequence repeats (SSRs) and five repeat units for all other SSRs. Larger repeats were examined using REPuter (Kurtz et al., 2001; Kurtz and Schleiermacher, 1999), using a minimum window size of 21 and a Hamming distance of 4.

Mega3 (Kumar et al., 2004) was used for some calculations, including GC content and codon usage, of the chloroplast genomes of *Pelargonium* and others as annotated in GenBank, including *Spinacia oleracea* (NC_002202), *Arabidopsis thaliana* (NC_000932), *Medicago truncatula* (NC_003119), *Lotus corniculatus* var. *japonicus* (NC_002694), and *Oenothera elata* subsp. *hookeri* (NC_002693).

Results

General Characteristics of the Genome. The chloroplast chromosome of *Pelargonium* × *hortorum* is the largest terrestrial plant chloroplast genome sequenced to date and can be represented as a circular molecule of 217,942 base pairs (bp) (Fig. 1; [Genbank Accession Number]). This is only slightly larger than previously estimated (Palmer et al., 1987). The genome has the stereotypical chloroplast quadripartite structure featuring two copies of a 75,741 bp IR separating a LSC region of 59,710 bp and a SSC region of 6,750 bp; these values are also very close to the 1987 estimates. In comparison with other genomes, these are about 3×, 2/3, and 1/3 of the usual sizes, respectively. For annotation purposes, the first base of the genome was defined as the first base of the LSC region where *trnH* is found and the plus or 'A' strand is designated as the strand on which *rbcL* is encoded. Approximately 46.8% of the genome encodes proteins, 1.4% encodes tRNAs, and 4.3% encodes ribosomal RNA. The non-coding regions (pseudogenes, spacers, and introns) account for the remaining 48.5% of the genome. GC content is 39.6% overall, 41.1% in coding regions, and 38.1% in non-coding regions. These GC values fall within the range of variation found for previously reported chloroplast genomes, and among the five genomes selected for direct comparison are

most similar to those for *Oenothera* (Table 1). Within the protein coding regions, both *Oenothera* and *Pelargonium* also share a similar pattern of codon usage, and generally have a slightly higher G+C content at all positions (Table 2).

Gene Content. Gene content is similar to that found in other angiosperm chloroplast genomes, although the total number of genes is dramatically higher due to duplications caused by massive IR expansion. The *Pelargonium* genome contains 76 unique protein genes (39 of which are duplicated within the IR, along with the first exon of *ndhA*), 4 rRNA genes (all of which are duplicated in the IR), and 29 tRNA genes (8 are duplicated in the IR, including *trnfM*-CAU, which has a third copy in the LSC). The total number of identified genes encoded is thus 161, with 51 genes duplicated within the IR (the usual complement is 15-20), and the average size of intergenic spacers is 368 bp.

Three genes found in tobacco cpDNA are apparently pseudogenes (*infA*, *ycf15*, and *rpoA*) and three others (*sprA*, *accD*, and *trnT*-GGU) are not detected at all in *Pelargonium*. *sprA* has been identified solely within the Solanaceae (Schmitz-Linneweber, 2002). The losses of *trnT*-GGU and *accD* both occur at rearrangement endpoints. The loss of this tRNA gene is not reflected in codon usage however, though it seems to be used at a uniformly low level in all the genomes examined (Table 2).

Only a few genes use alternate start codons. Both *rps19* and *rpl14* have a GTG start, which is a common feature of *rps19* in (angiosperm? land plant?) chloroplast genomes. The small photosystem gene *psbL* is also commonly annotated (in what group of plants???) as beginning with an ACG start. We infer that *cemA* and *ndhB* use ATT as a start codon, although alternatives are possible. In the former, the usual ATG start has been displaced by at least a 12 bp insertion if not a series of indels. In *ndhB*, the selection

of an ATT start seems clear, since there is an internal stop at the third codon from the usual start location. In the latter position, we also find an ACG codon which could serve as an initiator if C-to-T RNA editing changed this to ATG.

Small insertions and deletions (indels) relative to *Spinacia* are present in 32 genes, discounting length variation commonly seen at the 3′ terminus. The variable and large hypothetical coding frames *ycf1* (7,659 bp) and *ycf2* (6,333 bp) both have undergone numerous indel events, and while alignment of the former is nearly impossible outside of its terminal sequences, in the latter we estimate 48 indels ranging from 3-195 bp, although there are questionable alignments in several regions. The *P. hortorum ycf2* (*ORF2280*) sequence reported by Downie et al. (1994) is 99.8% identical to ours (10 nucleotide differences and three single base insertions that cause a temporary frame shift). Whether this is caused by cultivar differences or sequence errors is unclear. Other genes with multiple indels include the *23S* rRNA gene (five insertions of 4-95 bp, two deletions of 4-7 bp), *rpoB* (five insertions of 3-15 bp and three deletions of 3-9 bp), *rpoC1* (nine insertions of 3-18 bp), *rpoC2* (seven insertions of 3-21 bp and four deletions of 6-9 bp), and *rps18* (eight insertions of 3-27 bp). A 17-bp insertion induces a brief frame shift about 800 bp into *rpoC1*, but this is corrected six bp downstream by a one bp insertion.

Fifteen genes (for six tRNAs and nine proteins; eight of which are duplicated in the IR) contain a total of 17 introns, all of which maintain conserved intron boundaries. All introns but one are Group II self-splicing introns; the *trnL*-UAA intron is the sole Group I intron. A single intron has been lost from each of two ribosomal protein genes, *rps16* and

rpl16. The latter loss was noted previously in *Pelargonium* × *hortorum* and a species of *Erodium* (Campagna and Downie, 1998; Downie and Palmer, 1992).

Twenty-two pseudogenes were identified, originating from XX different genes. Four pseudogenes are found in the LSC, and the rest are within the IR. With the exception of *ΨinfA, Ψycf15,* and a family of *ΨrpoA* genes, which have no functional copies within the genome, and the two copies of *Ψrps14,* which are full copies of the functional gene interrupted by an internal stop codon, all of the identified pseudogenes are fragments of functional genes found elsewhere in the genome (*Ψrpl33, ΨtrnfM, Ψrrn16, Ψrpl23, ΨpetD, ΨrpoB,* and *ΨrpoC1*). Only *ΨinfA* is not part of the complex associations of repeated elements discussed in detail below.

As previously reported (Downie et al., 1994; Palmer et al., 1990a; Palmer et al., 1990b), a family of putative *rpoA* pseudogenes is located in the region surrounding *ycf2* in the IR (Fig. 2). Three fragments [α (ca. 650 bp), β (190 bp), and γ (415 bp)] with low identities to *rpoA* have been recognized. The β fragment is itself an extended partial repeat of the last 80 bp of the α fragment, and the γ fragment overlaps the β by 84 bp. The α and β fragments are the principal repeat subunits that characterize the repeats found in the second major repeat association (discussed below). Each of the three repeat units contains slightly different open reading frames (ORFs) (*ORF574*, *ORF332*, and *ORF365*), each containing both the α and β fragments, and thus these may represent a truncated *rpoA*-like gene. The γ fragment is also contained within a fourth ORF (*ORF221*). It is possible that one or all of these ORFs may retain functionality, but this was not determined in this study.

An additional ORF containing the IR-duplicated 5′ *ndhA* exon was designated as *ORF188*. Although fragmentary sequence identity with ORFs from other genomes was found, we have not annotated these features due to the lack of overall sequence and length conservation. However, the *trnA* intron contains two sequences with homology to previously recognized mitochondrial products in *Citrus* [*ACRS* (Ohtani et al., 2002)] and *Phaseolus* [*pvs-trnA* (Woloszynska et al., 2004)], and we have designated these as *ORF56* and *ORF42*, respectively.

Nucleotide Polymorphisms. Sequence polymorphisms were identified at eight locations based on having a minimum of two high quality sequence reads that deviated from the consensus sequence (Table 3). Two of these occur in intergenic spacers in the IR, and three others are non-synonymous changes found within protein-coding genes in the LSC. A single dinucleotide polymorphism was observed in the spacer between *rps16* and *trnQ*-UUG. Eleven length polymorphisms in mononucleotide simple sequence repeats were observed, two of which are duplicated in the IR. Only one of these falls within a coding region (*rps4*). Another one was originally thought to alter the coding frame for *ndhK* relative to that of tobacco, but this region is highly variable in comparison with other genomes, so an alternate start site was selected downstream of this variable region which appear not to be part of the gene.

Gene Order. In addition to its unusually large size, this genome is highly rearranged in comparison with the otherwise conserved order shared by tobacco and most other examined angiosperms (Palmer et al., 1987) (Fig. 1). The rearrangements include inversions, apparent translocations, deletions, and insertions of duplicated sequence. Considering only the order of genes and pseudogenes, over 34 rearrangements are

present, not including those duplications within the second copy of the IR (an additional 24). This is admittedly an artificially inflated number, as these rearrangements should be considered within the potentially simplifying context of an evolutionary hypothesis, but this serves to illustrate the overall complexity of this genome. However, evaluation of evolutionary hypotheses is a subjective process and several models that explain the data in different ways could be invoked (see Cosner et al., 1997). We attempt such interpretations later in our discussion.

On a local level, gene order is conserved within 25 blocks of genes (Fig. 1), and with one or two exceptions, all polycistronic operons appear to be preserved. However, These blocks are themselves highly rearranged in comparison with other angiosperm cpDNAs. These blocks range from about 30 bp to 30 kb, and contain from 1 to 25 genes or pseudogenes. It is also necessary to invoke three deletions to account for the loss of *trnT*, *accD*, and *ORF350* (the IR-duplicated portion of *ycf1* found between *ndhF* and *trnN* in tobacco). The largest blocks that appear in a similar relative arrangement and orientation to those of tobacco are two blocks within the LSC (blocks 1 and 8-9 in Fig. 1) and a block of SSC genes (block 24-25) (the contiguous blocks 8-9 and 24-25 are segregated due to the occurrence of 9 and 25 in the modern IR of *Pelargonium*). The IR is by far the most rearranged structural partition in the genome.

Another notable character of the genome that is associated with the rearrangements is the presence of the two repeat families noted by Palmer et al. (1987). These are located in 12 regions (only two are present in the LSC) and represent two different sets of associated repeats. These repeats are a complex series of duplications (i.e., insertions of duplicated sequence) that account for almost half of the number of rearrangements noted

above, as well as the majority of large repeats and all of the pseudogenes except *ΨinfA*. They are also associated with the potential disruption of one or possibly two operons.

The transcriptional linkage of the relatively short *rpl33-rps18* operon (blocks 12 and 13) is clearly disrupted by rearrangement. These two genes are neither associated with each other nor with their respective upstream or downstream partners as found in tobacco. Nonfunctional copies of *rpl33* also occur in several widely dispersed locations (see Fig.1) and are often associated with duplications of *rps14* and *trnfM*-CAU. These three genes are the characteristic components of the first repeat association (discussed in more detail below). These regions account for eight of the rearrangements noted above, although they have considerably more complex structure than this implies.

The *S10* or *rpl23* operon may also be disrupted at its terminus by the disruption of *rpoA*. As discussed earlier, the *rpoA* pseudogenes have a similar pattern of duplication to the genes discussed above and similarly constitute the major components of the second major association of repetitive elements. The latter also account for eight rearrangements, and thus the regions where these repeats occur represent high complexity hotspots containing about half of all rearrangements.

Relative to tobacco, the SSC is the least altered partition of *Pelargonium* (other than in size). Its only major changes are the translocations of *ndhF* (block 22 in fig.1) and *rpl32* (block 23) into different locations in the IR and the major expansion of the IR (block 26) to include all of *ycf1*, *rps15*, *ndhH* and part of *ndhA*.

Simple Sequence Repeats. We found a total of 440 exact or perfect microsatellite repeats within the *Pelargonium* genome (Table 4). The great majority of these (388) are 7-17 bp mononucleotide adenine or thymine runs, and slightly more than half of the latter

belong to the shortest class of only seven bp. Only six dinucleotide repeats of five units were found, and all of these are in the inverted repeat (three repeats with their complements). No other microsatellite types were detected.

Microsatellites are relatively evenly distributed throughout the genome (Fig. 1). Almost two thirds (280) are found within the IRs; the remaining third fall largely within the LSC region, with only 15 found in the SSC. Slightly more than half (245) occur within intergenic spacers, and roughly a third (157) occur in coding sequences. While introns represent only a small percentage of the genome's length, 38 SSRs are found within their boundaries, on average about two per intron.

Larger Repeats. Using REPuter, we further identified 6,698 repeats of 21 bp or larger with a sequence identity of greater than 80% within genome. The bulk (5,474, or 82%) are smaller repeats of 21-30 bp, and a large number of these are at least in part inexact mononucleotide SSRs that typically are interrupted by a transitional base or bases; many if not all of the previously discussed SSRs may be contained within this class. Despite the greater size of this genome, the number of repeats in this size class is remarkably uniform in comparisons with several other taxa for which genomic data are available (Fig. 3). However, this class represents 94% or more of the repeats in those other genomes. *Pelargonium* thus has a significantly larger number of 31 bp or larger repeats, having more than 3.6 times as many as *Oenothera*, and more than 35 times as many as *Spinacia*.

The sheer number of smaller repeats precludes a useful discussion of them here, so we focus here on the larger classes of 31 bp or more. Upon close examination, we found that 87% (1,065, including almost all of the largest class) of the larger repeats identified

are associated with the locations where Palmer et al. (1987) previously identified two families of repeats through hybridization studies. These are discussed in more detail below. The remaining 158 large repeats were ultimately reduced to nine pairs of dispersed repeats (31-104 bp) and six small, localized families of 15-33 bp tandem repeats with 4-12 repeats each (Table 1, Supplemental Data). Ten additional dispersed repeats (repeats *i-l*, *m-q*, Table 1, Supplemental Data) were also identified whose only other occurrence is in the repeat associations (see below) and their duplicates in the IR.

Analysis of this larger class provides some insight into how REPuter may overestimate repeat numbers. REPuter uses pairwise comparisons to recognize repeats, and this is the basis of the count; what is counted are the number of unique pairs, not the actual number of repeats. A repeat with multiple copies will thus be over-represented. REPuter may also compound this by recognizing several nested or overlapping series of repeats within a given region containing multiple repeats. For example, beginning in the 3' end of *rps*19, there is an 8-unit tandem repeat that extends 101 bp into the adjoining spacer. The basic repeat unit is 27 bp, with a degenerate unit of 21 bp. REPuter failed to identify the basic unit, and recognized 21 overlapping or nested repeats in this region. Similar situations are found in *ycf1*, *ycf2*, and the *5S-4.5S* spacer.

Repeat Families. In their study, Palmer et al. (1987) identified two families of dispersed repeats. We have identified these repeats in this sequence and verified that they fall into two major groups and confirmed that the mapping had properly placed them (Fig.1; Table 1, Supplemental Data). As noted earlier, almost half of the rearrangements in the genome and 87% of the larger repeats identified in the REPuter analysis are

localized in these regions; 25% of the smaller repeats (< 31 bp) fall here as well, as do all of the identified pseudogenes with the exception of *Ψinf*A.

Rather than simple families of repeats, however, these regions are composite assemblages of heterogeneous elements. A few unique elements (e.g., *rps18*) and 10 small dispersed repeat fragments from other regions of the genome are present, but most of the repeat elements are contained solely within these regions and are probably derived from "local" elements.

The first repeat association is the most complex, and is most readily recognized by the presence of *rpl33*, *trnfM*-CAU, or *rps14* and its respective pseudogenes (Fig. 4; also see Fig. 1). These repeats correspond to the nine-member family of Palmer et al. (1987), but rather than a single family this is an association of different repeat families. Members of this association occur in six locations (two in the LSC, and two duplicated in the IR; see Fig. 1; members 1.1-1.4, Table 1, supplemental data), each with its characteristic association of repetitive elements. Within these regions, we can further recognize several hierarchical classes of components. First, in addition to the gene families mentioned above, we can recognize a minimum of eight other major repeat elements (repeats *a-h*, Fig. 4c) unique to these regions and four small dispersed repeats (repeats *i-l*, Table 1, Supplemental Data) that represent small fragments (28-63 bp) of genes or spacers from diverse other parts of the genome. These elements are themselves subject to smaller scale (nucleotide) disruptions, duplications, insertions, deletions, and divergence and thus do not necessarily represent exact repeats. Secondly, while each of these elements appears at least once in a different context, certain arrangements of these elements are themselves repeated, and we refer to these as repeats *r1-r5* (Fig. 4c). Some unique sequence is also

present, mostly as small pieces of spacer with no discernable identity with anything else in the genome or in GenBank, and two non-repetitive genes (*trnG*-GCC, and *rps18*) are also found solely within these regions. The gene *rps18* is duplicated in the IR and represents the other half of the interupted *rpl33* operon. Whereas *rpl33* has been duplicated at least 4-7 times (depending on interpretation, and including the IR), *rps18* has not been subject to the same kind of duplication.

Percentage identity plots (Schwartz et al., 2003); Fig. 4a) of each member against the others illustrate the complexity of interpreting these repeats. The regions containing elements we identify as repeat elements *b* and *h* appear particularly subject to divergence, with many small duplications and low identities. Repeat *b* is duplicated 7 times, and is the most common element, absent only from member 1.4.

Overall similarity between the regions is illustrated in the pictorial alignments of figure 4b. These show that the shorter members 1.1 and 1.4 share many common elements, and that member 1.2 also shares these. Member 1.2 shares not only repeat elements, but also common arrangements of them, even within itself.

The second major repeat association is much simpler, and unlike the first repeat association, this has a more or less regular repeat structure of three units (members 2.1-2.3; Fig. 2) and thus can more properly be referred to as a family of repeats. This is the eight-member repeat family of Palmer et al. (1987), and these members are localized in the region of the IR surrounding *ycf2*. The arrangement of this region is similar to but not identical to that reported earlier for *P.* × *hortorum* (Downie et al., 1994; Palmer et al., 1990a; Palmer et al., 1990b), in which there appears to be an additional 3' fragment of *rpoA*. The basic repeat unit consists of three common repetitive elements, although

members 2.1 and 2.3 share three additional elements. The common elements include a ca. 120 bp sequence from the *rpoB-rpoC1* region of the LSC (repeat element *m*, Table 1, Supplemental Data) and the *rpoA* α and β pseudogenes. For convenience, we have designated the *rpoB-rpoC1* repeat fragment as *ΨrpoB/C1*, although it should be noted as two separate pseudogenes. This consists of the 3′ end of *rpoB*, the 5′ end of *rpoC1*, and the intervening spacer. Except for an 11 bp deletion in the *rpoB* segment, this repeat has 97% identity with its ancestral region in the LSC.

Members 2.1 and 2.3 are inverted copies and have an aligned sequence identity of 93%. They also share, as noted above, three repetitive elements not found in member 2.2: a 162 bp duplication of 3′ *rps11*, a 34 bp fragment with 88% identity to the *petB* intron (repeat *q*), and an 81-88 bp fragment with 95% identity to a piece of the *5S-4.5S* spacer (repeat *n*). A 45 bp fragment of the latter also follows *Ψrps11*, and thus member 2.3 is framed by two short direct repeats of this spacer region. Immediately upstream of member 2.3 in the *ycf2* spacer is a short, 37-bp fragment (repeat *o*) from a different region of the *5S-4.5S* spacer (95% identity). This is also in the opposite orientation relative to the two direct repeats.

Member 2.2 is inverted relative to 2.1, and is quite divergent, having a sequence identity of only 76%. The region between *ΨrpoC1/B* and *ΨrpoA* α is truncated and unalignable with the same region in the other members. The repeat as a whole is quite truncated in comparison, lacking the duplication of the *rps11* fragment as well as 800 bp of sequence that follows *ΨrpoA* β in repeats 2.1 and 2.3. This region is instead occupied by a fragment with a low amino acid identity (<40% with *Arabidopsis*) to 3′ *rpoA*; we designate this as *ΨrpoA*γ. This shares an 84 bp overlap with the β fragment. Together,

ΨrpoA α and γ represent a highly degenerate but nearly complete *rpoA*. Each of these is contained within a different ORF (*ORF332* and *ORF221*, respectively); these overlap slightly but are one base out of frame with each other. We investigated the possibility that an intron might have invaded *rpoA*. Although we identified potential splice sites, we were unable to fold a secondary structure that seemed consistent with either a Group II or degenerate Group III intron.

Discussion

General Characteristics. The chloroplast genome of *Pelargonium* × *hortorum* is remarkable for its overall size, inverted repeat size, number of rearrangements and repeats, and its apparent lack of a functional copy of *rpoA*. This study has largely confirmed the earlier estimates (Palmer et al., 1987) of overall size, structural partition sizes, placement of the LSC-IR boundaries, and the occurrence of two "families" of dispersed repeats, but has provided a much greater level of detail into the composition and structure of these repeats and the extent of gene order rearrangements.

Gene Content. Despite the vast increase in size of the genome, gene content is almost identical to that of other angiosperms. The loss of the tobacco *ORF350* can be ascribed to changes in the boundary of the IR and the complete duplication of *ycf1* rather than a true loss. Of the genes that have been lost, *accD* has also been lost in grasses (Katayama and Ogihara, 1996), Lobeliaceae (Knox and Palmer, 1999), and *Trachelium* (Campanulaceae; Cosner et al., 1997), and its loss here may be associated with its proximity to rearrangement endpoints. The loss of *trnT* also occurs at an inversion endpoint, and the presence of tRNAs has been often noted at those locations in grasses

(Hiratsuka et al., 1989; Howe et al., 1988; Shimada and Sugiura, 1989). With the exception of the wide-scale loss of tRNAs in *Epifagus* (Morden et al., 1991; Wolfe et al., 1992) and the related *Orobanche* (Lohan and Wolfe, 1998), tRNA loss seems to be rare within land plants; of the published land plant genomes, a single loss is reported, for *trnK* in *Adiantum* (Wolf et al., 2003).

Of the three genes present in the genome only as pseudogenes (*ΨinfA*, *Ψycf15* and *ΨrpoA*), only one, *infA* (translation initiation factor 1), has been previously reported as lost in a number of lineages, including *Pelargonium* (Millen et al., 2001). The potential lack of functionality of the other two pseudogenes may be open to question, however. The hypothetical gene *ycf15* is interrupted by a stop codon when compared to tobacco; similar results were reported for *Spinacia* (Schmitz-Linneweber, 2001), where it is annotated as a pseudogene. It remains to be determined whether this truncated product is transcribed and translated in *Pelargonium*, and thus we choose to be cautious in assigning functionality to this putative gene. Similarly, as noted earlier, the family of *rpoA* pseudogenes is contained within three ORFs, each having a conserved domain structure for an RNA polymerase alpha subunit. Any one or all of these could code for a truncated alpha subunit protein, but this was not determined in this study

While potential reading frames are quite numerous, we have chosen to conservatively note only those with sequence identity to genes of known function. In this assessment we found a great deal of conserved nucleotide sequence outside of recognized gene boundaries, and many of these regions have been previously characterized within various ORFs in other genomes. While we can identify strong sequence similarity on a local level, we rarely could find conservation over the full length of a potential reading frame.

For example, we found three different ORFs in the *trnI* intron that account for most of what has been identified as *ycf68* or *ORF133* in the grasses. These are out of frame with each other, however, and due to their fragmentary nature we decided not to recognize this feature. This seems to be the case even among closely related taxa [e.g., *Atropa* and *Nicotiana* (Schmitz-Linneweber, 2002))], and so caution seems advisable in the recognition of potential ORFs.

We have, however, noted two additional ORFs within this genome based on similarities to genes of known function. The first of these, *ORF56,* has also been identified in the chloroplast genome of *Calycanthus* (Goremykin et al., 2003). It is nearly identical (99%) to the mitochondrial ACR-toxin sensitivity (*ACRS*) gene of *Citrus jambhiri* Lush., and its presence has been noted in a number of chloroplast and mitochondrial genomes (Ohtani et al., 2002). The second ORF (*ORF42*) is a truncated 3′ fragment of another mitochondrial gene, *pvs-trnA* or *ORF98*, which is associated with a group of mitochondrial genes that impart cytoplasmic male sterility in a species complex of cultivated *Phaseolus* (Fabaceae) (Woloszynska et al., 2004). The situation of these two ORFs seems analogous to that of the many conserved sequences identified in our assessment of other ORFs, in that a BLAST search (Altschul et al., 1997) of GenBank reveals a large number of taxa with conserved chloroplast sequence of varying lengths and sequence identity. The lack of overall conservation across plant lineages suggests that while there may be some constraint on these sequences (e.g., constraints imposed by secondary structure of the intron), these ORFs probably do not represent functional genes in this genome, and it remains to be shown whether they are translated from the intron transcript.

Nucleotide Polymorphisms. Given that the genus *Pelargonium* is known to have biparental inheritance of plastids (Baur, 1909; James et al., 2001; Tilney-Bassett, 1973), it is remarkable that there are relatively few examples of heteroplasmy (but wasn't the DNA sequenced from multiple plants, so isn't plant to plant variation possible???) found in this study (Table 3), although this might be the result of varying patterns of inheritance (Tilney-Bassett and Amouslem, 1989; Tilney-Bassett and Birky, 1981). Most of the observed polymorphisms were present in low copy numbers relative to the consensus sequence, and while they could be the result of errors induced during PCR or sequencing, the reproducibility of these events specific to these locations would seem to point toward real polymorphism rather than artifact.

Gene Order, Repeats, and Repeat Associations. The size of the genome, its gene order, and the number and placement of repeats are all intimately connected. As inferred by Palmer et al. (1987), the increased size of the genome is largely due to gene duplication in the gross expansion of the IR, although the two repeat associations account for about 10% of the total length. While changes in the IR boundaries are common [the "ebb and flow" (Goulding et al., 1996; Price and Palmer, 1993)], large-scale changes are not. We can construct an evolutionary model in which a series of eight IR boundary shifts (a minimum of three contractions and five expansions) and six inversions (minimum) accounts for most of the major rearrangements (Fig. 5) found in the IR. Two small ebb-and-flow contractions (or a small and a large contraction) of the IR are all that is necessary to explain the placement of *trnI* at the beginning of the LSC, and a third can be invoked for the loss of the large ORF (*ORF350* in tobacco) representing the duplicated portion of *ycf1* between *ndhF* and *trnN*. Several waves of expansion can then be played

out that largely fit the current structure of the genome. These events explain the translocation of several conserved blocks of genes in the IR. Thus both large and small-scale changes in the IR boundaries have played an important role in restructuring gene order in *Pelargonium*.

It is possible that the IR could have been lost or severely reduced in size and content at some point. However, the necessary sequence of contractions and expansions seems to require the presence of both copies, at least until fairly late in the process when the composition and order of the IR was very much as it is today. While the large size of expansions and contractions suggested here might have been a series of smaller, ebb-and-flow events, we also see little evidence of this.

In addition to changes in the IR boundaries, inversions have played an important role in the evolution of the modern *Pelargonium* genome. In the simple model presented in Fig. 5, we hypothesize a minimum of only six inversions: 1) *psbD-ycf3* (blocks 3-7), 2) *psaI-rps18* (blocks 11-13), 3) re-inversion of *psbD-psbZ* (block 3), 4) re-inversion of *rps18* (block 11), 5) inversion of *ndhF-trnN* (blocks 20-21), and 6) 50 kb inversion of most of the newly expanded IR from *rpl20-trnN*. Upon re-examination of the data on the basis of this model, we discovered that inversions 3, 4, and 5 are each flanked by small inverted repeats (repeat 19, repeat element *l* of repeat member 1.4, and repeat 18, respectively, Table 1, supplemental data; the 24 bp inverted repeat that originally flanked *ndhF-trnN* has the appearance of a direct repeat due to the subsequent larger inversion.) We found no clear cases of such artifacts correlated with the other repeats, but analyses of these features is ongoing, and these could have been obscured either by sequence evolution or superimposition of other events. With the latter in mind, it is important to

note that inversions 1-4 are all adjacent to the locations of the first major repeat association, and important elements of those repeats were at least historically adjacent to or a part of these inversions. While Palmer et al. (1987) were unable to recognize that these repeats represented rearrangements themselves due to the limited resolution of filter hybridization, they had noted their placement near the ends of detected inversions and suggested recombination between them as the major cause of those inversions. Despite our failure to identify the small inverted repeats predicted to occur at all of these boundaries, this is still probable. The complexity of the repeats suggests that they have been subject themselves to a series of evolutionary events, and these could have obscured or eliminated signals of past events.

Our simple model of inversion and IR expansion shown in Fig. 5 does not account for the composition or arrangement of the repeat associations. These high complexity regions are a unique feature of this genome and account for many of the rearrangements present as well as the majority of the larger, non-microsatellite repeats detected. The two associations have no common elements, but do share a few common characteristics. Both are involved with the disruption and duplication of a gene or genes (in particular, *rpl33*, *rpoA*, *rps14* and *trnfM*) and at least potentially operons. Both contain a number of pseudogenes. Both involve elements that appear in novel combinations, and these combinations are duplicated and inverted. Many of the elements are endemic to the region of genome space in which they occur, but a few fragments from widely dispersed locations are present as well. The latter elements are typically drawn from otherwise non-repetitive regions without rearrangements.

The proximity of *rpoA* and *rpl33-rps18* at the ends of IR expansions suggests that these expansions possibly in conjunction with inversions could have disrupted their respective operons; similar situations are noted in *Trachelium* (Cosner et al., 1997) and *Vigna* (Perry et al., 2002). The repeat associations could be simply a record of the transcriptional recovery of functional genes lost in the breakup of these operons. Thus, while we cannot completely explain the complexity found in the repeat associations by these two processes, they may in fact have been the root causes of genomic instability that allowed these regions to evolve.

In this regard, the apparent lack of a functional *rpoA* is made more interesting by the fact that there are potentially three slightly different, shorter reading frames that could encode it. The transfer of *rpoA* to the nucleus and its subsequent loss in the chloroplast has been reported in mosses (Goffinet et al., 2005; Sugiura, 2003), and its loss has also been noted in the parasite *Cuscuta*, where it is related to the loss of photosynthesis (Krause et al., 2003). However, some evidence that a functional *rpoA* has been retained in the chloroplast of *Pelargonium* × *hortorum* has been suggested in other studies (Palmer et al., unpublished data; what/who should actually be cited here?).

The *rpoA* gene in several genera of Geraniaceae appears to be quite divergent (Mary Guisinger, pers. comm., Palmer et al.,?) and its functionality in this genome has been questioned (Ostrout and Kuhlman, 2003). The situation seems analogous to that of *Euglena*, where *rpoA* was not initially identified (Hallick et al., 1993), but was later found to be highly divergent and interrupted by the presence of an intron or introns (Sheveleva et al., 2002). The discovery of the *ΨrpoA*γ fragment was suggestive of the possibility that an intron had invaded the gene, and though we identified possible splice

sites, the brevity of the intervening sequence (about 340 bp) would have necessitated a highly reduced secondary structure. This was an exciting prospect, but we could not find a folding with even the reduced requirements of a Group III intron as found in *Euglena*.

Given that changes of the IR boundaries and inversions are the two major processes in the evolution of the genome, can we explain the complexity of the repeat regions by applying them? Possibly, but we need to invoke yet a third process. Much of our thinking about these high complexity regions could be simplified by the invasion of a duplicative transposable element or some mechanism that produces similar results. With the exception of the degenerate transposon in *Chlamydomonas* (Fan et al., 1995), tranposons are not known in plastids. An alternative explanation for the rampant duplication and inversion could be retroposition (Palmer, 1991). Retroposition (reverse transcription of an RNA transcript, in this case with the intron spliced out, to a cDNA, followed by recombination with the primary DNA sequence) has also been suggested as one method by which introns are lost (Bock et al., 1997; Dujon, 1989). Palmer (1991) notes that the presence of short dispersed pseudogene sequences may support the idea of random incorporation of cDNAs. Such a process could account for the seemingly random incorporation of non-regionally endemic DNA into the hotspot regions, but not why the more endemic elements (e.g., *rpl33*) are themselves repeated so often. Given the nature of these repeat associations, it is very likely that they are subject to both intra- and intermolecular recombination, and this could also result in duplications (Howe et al., 1988).

In figures 6 and 7, we extend the simple model of evolution presented in figure 5 to the special cases involving the two repeat associations by adding putatively ancestral

duplications. In each of these models, we make two simplifying assumptions. First, we assume that duplications occurred prior to any other rearrangements (i.e., inversions and IR shifts) that directly involve the duplicated elements, and second that these are not just simple tandem duplications of a single gene, but involve various duplications of one or several elements. Evidence for the latter is that both *rps14* and *trnfM* are duplicated in two different putatively ancestral arrangements. Once these duplications were in place, then a relatively simple series of seesaw-like inversions and IR boundary displacements, some of which create orphan fragments, could account for almost all of the current structure we see in these regions. In combining all of these evolutionary models, a total of eight IR shifts, 12 inversions, and eight duplications are required at a minimum to explain the structure of the modern *Pelargonium* genome.

If our assumption of the temporal??? priority of duplications is correct, then it may be that duplications involving *rpoA* and *rpl33* could have interrupted their respective transcriptional operons rather than the processes of inversion and IR shifts mentioned earlier. Similarly, duplication of *rps14* may have disrupted its operon as well, and thus this may be the root cause of genomic instability that resulted in numerous inversions and IR boundary shifts.

Understanding of the processes involved in the evolution of these highly complex regions will require the continued close examination of the smaller repeats, as well as the sequencing of several closely related genomes with fewer rearrangements. While the number of repeats based on the REPuter analysis may be greatly exaggerated, there seems to be a previously undocumented presence of many repeats of less than 30 bp in all genomes examined, and despite the numeric susceptibilities it is not clear either of them

appears to be more or less uniform despite differences in size, structure and content. A cursory examination reveals that many of these lesser repeats consist of imperfect SSRs or combinations of SSRs, and this could be a background of evolutionary noise. However, preliminary analysis shows that similarly structured repeats do seem to play a role in rearrangements with inversions and possibly in changes of the IR (Goulding et al., 1996). Given this background level of repeats, the question might not be why is *Pelargonium* so highly rearranged, but why aren't rearrangements more common in all chloroplast genomes?

In summary, the chloroplast genome of *Pelargonium* × *hortorum* is both the largest and most rearranged genome yet sequenced among land plants. The large increase in size and the number of rearrangements are correlated with a series of large expansions of the inverted repeat and inversions. These may have resulted in the disruption of transcriptional operons, and genes involved in these disruptions form the core units of a series of large, complex repeats that are unique characters of this genome. These repeat regions are hotspots for duplications, duplicated inversions, and the incorporation of a few other repetitive elements from elsewhere in the genome. In addition to the two major processes of inversion and large shifts in IR boundaries, a process of sequence duplication may be at work, possibly including the invasion of transposons, a relatively regular process of retroposition, and/or frequent recombination. Despite the major increase in size and complexity, the gene content of this genome is similar to that of other angiosperms. Exceptions to this are the large number of pseudogenes associated with large repeats, the recognition of two ORFs in the *trnA* intron previously identified from mitochondrial genomes, and in particular, the lack of a recognizably functional *rpoA*.

Acknowledgements

This work was supported by grant DEB-0120709 from the National Science Foundation. Part of this work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, Lawrence Berkeley National Laboratory, under contract No. DE-AC02-05CH11231

Literature Cited

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. Journal of Molecular Biology 215:403-410.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25:3389-3402.

Baur, E. 1909. Das Wesen und die Erblichkeitsverhältnisse der "Varietates albomarginatae hort." von "*Pelargonium zonale"*. Zeitschr.f. Indukt. Abstammungs und Vererbungslehre 1:330-351.

Bock, R., M. Hermann, and M. Fuchs. 1997. Identification of critical nucleotide positions for plastid RNA editing site recognition. Rna-a Publication of the Rna Society 3:1194-1200.

Bruneau, A., J. J. Doyle, and J. D. Palmer. 1990. A chloroplast DNA inversion as a subtribal character in the Phaseoleae (Leguminosae). Systematic Botany 15:378-386.

Campagna, M. L., and S. R. Downie. 1998. The intron in chloroplast gene *rpl16* is missing from the flowering plant families Geraniaceae, Goodeniaceae, and Plumbaginaceae. Transactions of the Illinois State Academy of Science 91:1-11.

Cosner, M. E., R. K. Jansen, J. D. Palmer, and S. R. Downie. 1997. The highly rearranged chloroplast genome of *Trachelium caeruleum* (Campanulaceae): Multiple inversions, inverted repeat expansion and contraction, transposition, insertions/deletions, and several repeat families. Current Genetics 31:419-429.

Cosner, M. E., L. A. Raubeson, and R. K. Jansen. 2004. Chloroplast DNA rearrangements in Campanulaceae: phylogenetic utility of highly rearranged genomes. BMC Evolutionary Biology 4:1-17.

Downie, S. R., D. S. Katzdownie, K. H. Wolfe, P. J. Calie, and J. D. Palmer. 1994. Structure and evolution of the largest chloroplast gene (ORF2280) - internal plasticity and multiple gene loss during angiosperm evolution. Current Genetics 25:367-378.

Downie, S. R., and J. D. Palmer. 1992. Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. Pages 14-35 *in* Molecular Systematics of Plants (P. S. Soltis, D. E. Soltis, and J. J. Doyle, eds.). Chapman and Hall, New York.

Doyle, J. J., J. I. Davis, R. J. Soreng, D. Garvin, and M. J. Anderson. 1992. Chloroplast DNA inversions and the origin of the grass family (Poaceae). Proceedings of the National Academy of Sciences of the United States of America 89:7722-7726.

Doyle, J. J., J. L. Doyle, J. A. Ballenger, and J. D. Palmer. 1996. The distribution and phylogenetic significance of a 50-kb chloroplast DNA inversion in the flowering plant family Leguminosae. Molecular Phylogenetics and Evolution 5:429-438.

Dujon, B. 1989. Group I introns as mobile genetic elements: facts and mechanisitc speculations - a review. Gene 82:91-114.

Ewing, B., and P. Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research 8:186-194.

Ewing, B., L. Hillier, M. C. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research 8:175-185.

Fan, W. H., M. A. Woelfle, and G. Mosig. 1995. 2 copies of a DNA element, Wendy, in the chloroplast chromosome of *Chlamydomonas reinhardtii* between rearranged gene clusters. Plant Molecular Biology 29:63-80.

Goffinet, B., N. J. Wickett, A. J. Shaw, and C. J. Cox. 2005. Phylogenetic significance of the *rpoA* loss in the chloroplast genome of mosses. Taxon 54:353-360.

Gordon, D., C. Abajian, and P. Green. 1998. Consed: A graphical tool for sequence finishing. Genome Research 8:195-202.

Goremykin, V., K. I. Hirsch-Ernst, S. Wolfl, and F. H. Hellwig. 2003. The chloroplast genome of the "basal" angiosperm *Calycanthus fertilis* - structural and phylogenetic analyses. Plant Systematics and Evolution 242:119-135.

Goulding, S. E., R. G. Olmstead, C. W. Morden, and K. H. Wolfe. 1996. Ebb and flow of the chloroplast inverted repeat. Molecular & General Genetics 252:195-206.

Hachtel, W., A. Neuss, and J. Vomstein. 1991. A chloroplast DNA inversion marks an evolutionary split in the genus *Oenothera*. Evolution 45:1050-1052.

Hallick, R. B., L. Hong, R. G. Drager, M. R. Favreau, A. Monfort, B. Orsat, A. Spielmann, and E. Stutz. 1993. Complete sequence of *Euglena gracilis* chloroplast DNA. Nucleic Acids Research 21:3537-3544.

Hiratsuka, J., H. Shimada, R. Whittier, T. Ishibashi, M. Sakamoto, M. Mori, C. Kondo, Y. Honji, C. R. Sun, B. Y. Meng, Y. Q. Li, A. Kanno, Y. Nishizawa, A. Hirai, K. Shinozaki, and M. Sugiura. 1989. The complete sequence of the rice (*Oryza sativa*) chloroplast genome - intermolecular recombination between distinct transfer RNA Genes accounts for a major plastid DNA inversion during the evolution of the cereals. Molecular & General Genetics 217:185-194.

Hoot, S. B., and J. D. Palmer. 1994. Structural rearrangements, including parallel inversions, within the chloroplast genome of *Anemone* and related genera. Journal of Molecular Evolution 38:274-281.

Howe, C. J., R. F. Barker, C. M. Bowman, and T. A. Dyer. 1988. Common features of 3 inversions in wheat chloroplast DNA. Current Genetics 13:343-349.

Hupfer, H., Swaitek, M., Hornung, S., Herrmann, R. G., Maier, R. M., Chiu, W.-L. and Sears, B. 2000. Complete nucleotide sequence of the *Oenothera elata* plastid chromosome, representing plastome 1 of the five distinguishable *Euoenthera* plastomes. Mol Gen Genet 263:581-585.

James, C. M., J. A. Barrett, S. J. Russell, and M. Gibby. 2001. A rapid PCR based method to establish the potential for paternal inheritance of chloroplasts in *Pelargonium*. Plant Molecular Biology Reporter 19:163-167.

Jansen, R. K., and J. D. Palmer. 1987. A chloroplast DNA inversion marks an ancient evolutionary split in the sunflower family (Asteraceae). Proceedings of the National Academy of Sciences of the United States of America 84:5818-5822.

Jansen, R. K., L. A. Raubeson, J. L. Boore, C. W. dePamphilis, T. W. Chumley, R. C. Haberle, S. K. Wyman, A. Alverson, R. Peery, S. J. Herman, H. M. Fourcade, J. V. Kuehl, J. R. McNeal, J. Leebens-Mack, and L. Cui. 2005. Methods for obtaining and analyzing whole chloroplast genome sequences. Pages 348-383 *in* Molecular Evolution, Producing the Biochemical Data, Part B (E. A. Zimmer, and E. Roalson, eds.). Academic Press.

Johansson, J. T. 1999. There large inversions in the chloroplast genomes and one loss of the chloroplast gene *rps16* suggest an early evolutionary split in the genus *Adonis*

(*Ranunculaceae*) (vol 218, pg 133, 1999). Plant Systematics and Evolution 218:318-318.

Johansson, J. T., and R. K. Jansen. 1991. Chloroplast DNA variation among 5 species of Ranunculaceae - structure, sequence divergence, and phylogenetic relationships. Plant Systematics and Evolution 178:9-25.

Johansson, J. T., and R. K. Jansen. 1993. Chloroplast DNA variation and phylogeny of the Ranunculaceae. Plant Systematics and Evolution 187:29-49.

Katayama, H., and Y. Ogihara. 1996. Phylogenetic affinities of the grasses to other monocots as revealed by molecular analysis of chloroplast DNA. Current Genetics 29:572-581.

Kim, K. J., K. S. Choi, and R. K. Jansen. 2005. Two chloroplast DNA inversions originated simultaneously during early evolution in the sunflower family. Molecular Biology and Evolution 22:1783-1792.

Kim, Y. D., and R. K. Jansen. 1994. Characterization and phylogenetic distribution of a chloroplast DNA rearrangement in the Berberidaceae. Plant Systematics and Evolution 193:107-114.

Knox, E. B., S. R. Downie, and J. D. Palmer. 1993. Chloroplast genome rearrangements and the evolution of giant *Lobelia*s from herbaceous ancestors. Molecular Biology and Evolution 10:414-430.

Knox, E. B., and J. D. Palmer. 1999. The chloroplast genome arrangement of *Lobelia thuliniana* (Lobeliaceae): expansion of the inverted repeat in an ancestor of the Campanulales. Plant Systematics and Evolution 214:49-64.

Kolodner, R., and K. K. Tewari. 1972. Molecular size and conformation of chloroplast deoxyribonucleic acid from pea leaves. Journal of Biological Chemistry 247:6355–6364.

Krause, K., S. Berg, and K. Krupinska. 2003. Plastid transcription in the holoparasitic plant genus *Cuscuta*: Parallel loss of the *rrn16* PEP-promoter and of the *rpoA* and *rpoB* genes coding for the plastid-encoded RNA polymerase. Planta 216:815-823.

Kumar, S., K. Tamura, and M. Nei. 2004. MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Briefings in Bioinformatics 5:150-163.

Kurtz, S., J. V. Choudhuri, E. Ohlebusch, C. Schleiermacher, J. Stoye, and R. Giegerich. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Research 29:4633-4642.

Kurtz, S., and C. Schleiermacher. 1999. REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics 15:426-427.

Lidholm, J., A. E. Szmidt, J.-E. Hallgren, and P. Gustafsson. 1988. The chloroplast genomes of conifers lack one of the rDNA-encoding inverted repeats. Molecular Genetics and Genomics 212:6-10.

Lohan, A. J., and K. H. Wolfe. 1998. A subset of conserved tRNA genes in plastid DNA of nongreen plants. Genetics 150:425-433.

Millen, R. S., R. G. Olmstead, K. L. Adams, J. D. Palmer, N. T. Lao, L. Heggie, T. A. Kavanagh, J. M. Hibberd, J. C. Giray, C. W. Morden, P. J. Calie, L. S. Jermiin, and K. H. Wolfe. 2001. Many parallel losses of *infA* from chloroplast DNA

during angiosperm evolution with multiple independent transfers to the nucleus. Plant Cell 13:645-658.

Milligan, B. G., J. N. Hampton, and J. D. Palmer. 1989. Dispersed repeats and structural reorganization in subclover chloroplast DNA. Molecular Biology and Evolution 6:355-368.

Morden, C. W., K. H. Wolfe, C. W. Depamphilis, and J. D. Palmer. 1991. Plastid Translation and Transcription Genes in a Nonphotosynthetic Plant - Intact, Missing and Pseudo Genes. Embo Journal 10:3281-3288.

Ohtani, K., H. Yamamoto, and K. Akimitsu. 2002. Sensitivity to *Alternaria alternata* toxin in citrus because of altered mitochondrial RNA processing. Proceedings of the National Academy of Sciences of the United States of America 99:2439-2444.

Ostrout, N. D., and P. L. Kuhlman. Year. Development of an *in vitro* expression system for the *rpo* genes found within the *Pelargonium* chloroplast genome *in* FASEB Meeting on Experimental Biology: Translating the Genome, San Diego, CA, USA. 17:Abstract 371.1.

Palmer, J. D. 1990. Contrasting Modes and Tempos of Genome Evolution in Land Plant Organelles. Trends in Genetics 6:115-120.

Palmer, J. D. 1991. Plastid chromosomes: structure and evolution. Pages 5-53 *in* Molecular Biology of Plastids (L. Bogorad, ed.) Academic Press, Orlando, FL.

Palmer, J. D., S. L. Baldauf, P. J. Calie, and C. W. dePamphilis. 1990a. Chloroplast gene instability and transfer to the nucleus. Pages 97-106 *in* Molecular Evolution (M. T. Clegg, and S. J. O'Brien, eds.). Alan R. Liss, Inc., New York.

Palmer, J. D., P. J. Calie, C. W. dePamphilis, J. M. J. Logsdon, D. S. Katz-Downie, and S. R. Downie. 1990b. An evolutionary genetic approach to understanding plastid gene function: lessons from photosynthetic and nonphotosynthetic plants. Pages 475-482 *in* Current Research in Photosynthesis (M. Baltscheffsky, ed.) Kluwer Academic Publishers, Amsterdam.

Palmer, J. D., J. M. Nugent, and L. A. Herbon. 1987. Unusual structure of geranium chloroplast DNA - A triple-sized inverted repeat, extensive gene duplications, multiple inversions, and 2 repeat families. Proceedings of the National Academy of Sciences of the United States of America 84:769-773.

Palmer, J. D., B. Osorio, and W. F. Thompson. 1988. Evolutionary significance of inversions in legume chloroplast DNAs. Current Genetics 14:65-74.

Palmer, J. D., and W. F. Thompson. 1981. Rearrangements in the chloroplast genomes of mung bean and pea. Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences 78:5533-5537.

Perry, A. S., S. Brennan, D. J. Murphy, T. A. Kavanagh, and K. H. Wolfe. 2002. Evolutionary re-organisation of a large operon in adzuki bean chloroplast DNA caused by inverted repeat movement. DNA Research 9:157-162.

Price, R. A., and J. D. Palmer. 1993. Phylogenetic relationships of the Geraniaceae and Geraniales from rbcL sequence comparisons. Annals of the Missouri Botanical Garden 80:661-671.

Raubeson, L. A., and R. K. Jansen. 1992. A rare chloroplast DNA structural mutation is shared by all conifers. Biochemical Systematics and Ecology 20:17-24.

Raubeson, L. A., and R. K. Jansen. 2005. Chloroplast genomes of plants. Pages 45-68 *in* Plant Diversity and Evolution: Genotypic and Phenotypic Variation in Higher Plants (R. J. Henry, ed.) CAB International, Cambridge, MA.

Schmitz-Linneweber, C., Maier, R, M., Alcaraz, J-P., Cottet, A., Herrmann R. G. and Mache, R. 2001. The plastid chromosome of spinach (*Spinacia oleracea*): complete nucleotide sequence and gene organization. Plant Molecular Biology 45:307-315.

Schmitz-Linneweber, C., Rege, R., Du, T, G., Hupfer, H., Herrmann, R, G. and Maier, R, M. 2002. The plastid chromosome of *Atropa belladonna* and its comparison with that of *Nicotiana Tabacum*: The role of RNA editing in generating divergence in the process of plant speciation. mol. Biol. Evol. 19:1602-1612.

Schwartz, S., L. Elnitski, M. Li, M. Weirauch, C. Riemer, A. Smit, N. C. S. Program, E. D. Green, R. C. Hardison, and W. Miller. 2003. MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences. Nucleic Acids Research 31:3518-3524.

Sheveleva, E. V., N. V. Giordani, and R. B. Hallick. 2002. Identification and comparative analysis of the chloroplast alpha-subunit gene of DNA-dependent RNA polymerase from seven *Euglena* species. Nucleic Acids Research 30:1247-1254.

Shimada, H., and M. Sugiura. 1989. Pseudogenes and short repeated sequences in the rice chloroplast genome. Current Genetics 16:293-301.

Shinozaki, K., M. Ohme, M. Tanaka, T. Wakasugi, N. Hayashida, T. Matsubayashi, N. Zaita, J. Chunwongse, J. Obokata, K. Yamaguchishinozaki, C. Ohto, K. Torazawa, B. Y. Meng, M. Sugita, H. Deno, T. Kamogashira, K. Yamada, J.

Kusuda, F. Takaiwa, A. Kato, N. Tohdoh, H. Shimada, and M. Sugiura. 1986. The complete nucleotide sequence of the tobacco chloroplast genome - its gene organization and expression. Embo Journal 5:2043-2049.

Strauss, S. H., J. D. Palmer, G. T. Howe, and A. H. Doerksen. 1988. Chloroplast genomes of 2 conifers lack a large inverted repeat and are extensively rearranged. Proceedings of the National Academy of Sciences of the United States of America 85:3898-3902.

Sugiura, C., Kobayashi Y., Aoki, S., Sugita, C. and Sugita M. 2003. Complete chloroplast DNA sequence of the moss *Physcomitrella patens*: evidence for the loss and relocation of *rpoA* from the chloroplast to the nucleus. Nucleic Acids Research 31:5324-5331.

Thurston, M. I., and D. Field. 2005. Msatfinder: detection and characterisation of microsatellites, version 1.6.8.

Tilney-Bassett, R. A. E. 1973. The control of plastid inheritance in *Pelargonium*. II. Heredity 30:1-13.

Tilney-Bassett, R. A. E., and A. B. Amouslem. 1989. Variation in plastid inheritance between *Pelargonium* cultivars and their hybrids. Heredity 63:145-153.

Tilney-Bassett, R. A. E., and C. W. Birky, Jr. 1981. The mechanism of the mixed inheritance of chloroplast genes in *Pelargonium*: Evidence from gene frequency distributions among the progeny of crosses. Theoretical and Applied Genetics 60:43-53.

Wakasugi, T., J. Tsudzuki, T. Ito, K. Nakashima, T. Tsudzuki, and S. M. 1994. Loss of all *ndh* genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. Proc. Natl. Acad. Sci. USA 91:9794-9798.

Wolf, P. G., C. A. Rowe, R. B. Sinclair, and M. Hasebe. 2003. Complete nucleotide sequence of the chloroplast genome from a leptosporangiate fern, *Adiantum capillus-veneris* L. DNA Research 10:59-65.

Wolfe, K. H., C. W. Morden, and J. D. Palmer. 1992. Function and Evolution of a Minimal Plastid Genome from a Nonphotosynthetic Parasitic Plant. Proceedings of the National Academy of Sciences of the United States of America 89:10648-10652.

Woloszynska, M., T. Bocer, P. Mackiewicz, and H. Janska. 2004. A fragment of chloroplast DNA was transferred horizontally, probably from non-eudicots, to mitochondrial genome of *Phaseolus*. Plant Molecular Biology 56:811-820.

Wyman, S. K., R. K. Jansen, and J. L. Boore. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252-3255.

Table 1. Nucleotide composition

Taxon	T(U)	C	A	G
Total				
Spinacia	31.9	18.8	31.3	18
Arabidopsis	32.3	18.4	31.4	17.9
Medicago	33.1	17.6	32.9	16.4
Lotus	32.1	17.9	31.9	18.1
Oenothera	30.8	19.5	30.1	19.6
Pelargonium	30.3	19.9	30.1	19.7
Coding				
Spinacia	31.8	17.4	30.7	20.1
Arabidopsis	32	17.3	30.8	19.9
Medicago	32.3	16.8	31.1	19.8
Lotus	32	17.1	31.1	19.8
Oenothera	30.3	18.4	30	21.3
Pelargonium	30.6	18.5	29.8	21.1
1st position				
Spinacia	23.7	18.7	30.2	27.4
Arabidopsis	23.9	18.5	30.5	27.1
Medicago	23.7	18.2	30.8	27.3
Lotus	23.8	18.2	31.1	26.9
Oenothera	22.6	19.3	29.9	28.2
Pelargonium	23.1	19.7	29.8	27.4
2nd position				
Spinacia	33	20.4	28.9	17.8
Arabidopsis	33.2	20.2	28.9	17.7
Medicago	33.4	20.1	29.2	17.3
Lotus	33.4	20.1	29.2	17.3
Oenothera	32	20.7	29	18.3
Pelargonium	32.4	20.5	29	18.1
3rd position				
Spinacia	38.7	13	33	15.2
Arabidopsis	39	13	33	14.9

Medicago	39.7	12.2	33.3	14.8
Lotus	39	13	32.9	15.2
Oenothera	36.2	15.3	31.1	17.4
Pelargonium	36.2	15.5	30.6	17.7

Table 2. Codon usage

codon		*Spinacia*	*Arabidopsis*	*Medicago*	*Lotus*	*Oenothera*	*Pelargonium*
UUU	(F)	0.041	0.042	0.043	0.042	0.035	0.038
UUC	(F)	0.018	0.019	0.016	0.018	0.020	0.020
UUA	(L)	0.037	0.037	0.036	0.035	0.032	0.031
UUG	(L)	0.020	0.019	0.022	0.021	0.021	0.022
UCU	(S)	0.021	0.021	0.021	0.022	0.020	0.020
UCC	(S)	0.011	0.011	0.011	0.011	0.012	0.012
UCA	(S)	0.015	0.015	0.015	0.015	0.013	0.013
UCG	(S)	0.006	0.007	0.006	0.008	0.008	0.007
UAU	(Y)	0.030	0.030	0.032	0.030	0.027	0.028
UAC	(Y)	0.007	0.006	0.006	0.006	0.008	0.007
UAA	(*)	-	-	-	-	-	0.002
UAG	(*)	-	-	-	-	-	0.001
UGU	(C)	0.008	0.009	0.008	0.009	0.008	0.007
UGC	(C)	0.003	0.003	0.002	0.003	0.003	0.003
UGA	(*)	-	-	-	-	-	0.001
UGG	(W)	0.018	0.018	0.018	0.018	0.018	0.018
CUU	(L)	0.022	0.022	0.022	0.021	0.022	0.027
CUC	(L)	0.006	0.007	0.006	0.006	0.008	0.008
CUA	(L)	0.015	0.014	0.015	0.014	0.014	0.015
CUG	(L)	0.006	0.006	0.006	0.006	0.008	0.008
CCU	(P)	0.018	0.016	0.017	0.015	0.015	0.017
CCC	(P)	0.007	0.007	0.007	0.009	0.010	0.009
CCA	(P)	0.012	0.012	0.013	0.011	0.011	0.011
CCG	(P)	0.005	0.005	0.005	0.005	0.007	0.007
CAU	(H)	0.017	0.017	0.018	0.017	0.016	0.016
CAC	(H)	0.006	0.006	0.005	0.005	0.006	0.007
CAA	(Q)	0.029	0.029	0.027	0.029	0.028	0.026
CAG	(Q)	0.007	0.008	0.007	0.008	0.009	0.009
CGU	(R)	0.014	0.013	0.013	0.013	0.014	0.014
CGC	(R)	0.004	0.005	0.004	0.004	0.005	0.006
CGA	(R)	0.014	0.014	0.014	0.014	0.014	0.013

CGG	(R)	0.004	0.004	0.004	0.004	0.005	0.005
AUU	(I)	0.043	0.044	0.045	0.045	0.039	0.037
AUC	(I)	0.015	0.015	0.015	0.016	0.017	0.017
AUA	(I)	0.027	0.027	0.029	0.029	0.024	0.023
AUG	(M)	0.023	0.022	0.022	0.022	0.022	0.022
ACU	(T)	0.021	0.021	0.021	0.020	0.019	0.020
ACC	(T)	0.009	0.009	0.009	0.010	0.010	0.010
ACA	(T)	0.016	0.016	0.016	0.016	0.014	0.014
ACG	(T)	0.005	0.005	0.005	0.005	0.006	0.006
AAU	(N)	0.036	0.036	0.038	0.037	0.033	0.029
AAC	(N)	0.011	0.011	0.010	0.011	0.011	0.011
AAA	(K)	0.042	0.044	0.045	0.045	0.042	0.043
AAG	(K)	0.012	0.011	0.011	0.013	0.014	0.019
AGU	(S)	0.015	0.015	0.015	0.015	0.014	0.013
AGC	(S)	0.005	0.004	0.004	0.004	0.006	0.006
AGA	(R)	0.017	0.017	0.017	0.017	0.018	0.018
AGG	(R)	0.006	0.006	0.006	0.006	0.008	0.008
GUU	(V)	0.021	0.021	0.022	0.021	0.020	0.020
GUC	(V)	0.006	0.006	0.006	0.006	0.008	0.009
GUA	(V)	0.021	0.020	0.022	0.022	0.020	0.019
GUG	(V)	0.007	0.008	0.007	0.007	0.009	0.008
GCU	(A)	0.026	0.027	0.027	0.025	0.025	0.026
GCC	(A)	0.009	0.008	0.008	0.009	0.011	0.010
GCA	(A)	0.016	0.015	0.016	0.015	0.015	0.014
GCG	(A)	0.007	0.006	0.006	0.006	0.009	0.008
GAU	(D)	0.032	0.031	0.032	0.032	0.031	0.028
GAC	(D)	0.008	0.007	0.007	0.007	0.010	0.010
GAA	(E)	0.041	0.041	0.041	0.039	0.039	0.038
GAG	(E)	0.012	0.012	0.012	0.012	0.015	0.016
GGU	(G)	0.023	0.023	0.024	0.023	0.022	0.021
GGC	(G)	0.007	0.007	0.005	0.006	0.009	0.009
GGA	(G)	0.028	0.028	0.028	0.027	0.025	0.024
GGG	(G)	0.012	0.011	0.010	0.011	0.014	0.014

Table 3. Observed Polymorphisms

Polymorphism type	start	end	location	description
length	4732	4744	trnK intron	poly-T, 14 or 15
dinucleotide	6053	6054	IGS rps16-trnQ	GA or TC
single nucleotide	7066		IGS psbK-psbI	A or T
length	7936	7948	IGS trnS-trnG	poly-A, 13 or 12
single nucleotide	8005		IGS trnS-trnG	G or T
single nucleotide	16173		rpoC2	T or C; silent, 3rd position
single nucleotide	26397		rpoB	A or C; first position, V->G
length	30630	30644	rpl33 pseudogene	poly-A, 15 or 16
single nucleotide	40473		IGS rps14-rps14 pseudogene	C or T
length	50342	50345	rps4	poly-A, 4 or 5; frameshift mutation
length	51469	51480	IGS trnT-trnL	poly-T, 12 or 10
length	53332	53343	IGS trnF-ndhJ	poly-G, 12 or undetermined
single nucleotide	58826		atpB	C or T; silent, 3rd position, M->I
single nucleotide	59318		IGS atpB-rbcL	A or C
length	64349	64362	IGS 16S pseudogene-trnfM	poly-T, 14 or undetermined
single nucleotide	103722		IGS rpoAa2 pseudogene - rpoB/C1 fragment	A or G
single nucleotide	103808		rpoB/C1 fragment	T or G
length	115570	115584	IGS petD-petB	poly-A, 15 or 14
length	138542	138554	IGS ndhD-psaC	poly-T, 13 or undetermined
length	162071	162085	IGS petD-petB	poly-T, 15 or 14; reverse complement of 115570
single nucleotide	173847		rpoB/C1 fragment	A or C; reverse complement of 103808
single nucleotide	173933		IGS rpoAa2 pseudogene - rpoB/C1 fragment	T or C; reverse complement of 103722

length	213293	213306	IGS rbcL-LSC (trnI)	poly-A, 14 or undetermined; reverse complement of 64349

Table 4. Comparison of SSRs

motif	Spinacia	Arabidopsis	Medicago	Lotus	Oenothera	Pelargonium
a	150	204	161	208	135	199
c	9	17	7	13	18	22
g	4	13	1	11	17	24
t	153	253	186	208	144	188
ac	1	0	0	0	0	0
ag	0	0	0	0	0	1
at	8	18	26	27	6	4
ct	0	0	2	0	0	1
aat	0	1	0	1	0	0
act	0	0	1	0	0	0
att	0	1	0	1	0	0
SUMMARY						
mononucleotide	316	487	355	440	314	434
dinucleotide	9	18	28	27	6	6
trinucleotide	0	2	1	2	0	0
Total	325	507	384	469	320	440

Figure Legends

Fig. 1. Map of the chloroplast genome of *Pelargonium* × *hortorum* L. H. Bailey. The genome consists of 217,942 bp and features a large single copy region (LSC) of 59,710 bp, a small single copy region (SSC) of 6,750 bp, and two copies of an inverted repeat (IR) of 75,741 bp. Middle ring shows the locations of exact SSRs (small hash marks), larger repeats (large hash marks), and the two major repeat associations (1.1-1.4; 2.1-2.3). Interior ring details rearrangements with blocks of genes numbered in the order in which they appear in tobacco; inversions are shaded.

Fig. 2. Repeat association or family II and associated elements surrounding the *ycf2* region of the inverted repeat. Black bars indicate the repeats.

Fig. 3. Histogram of repeat size frequency in *Pelargonium* and five related genomes. Repeat size classes are 21-30 bp; 31-50 bp; 51-100 bp; > 100 bp.

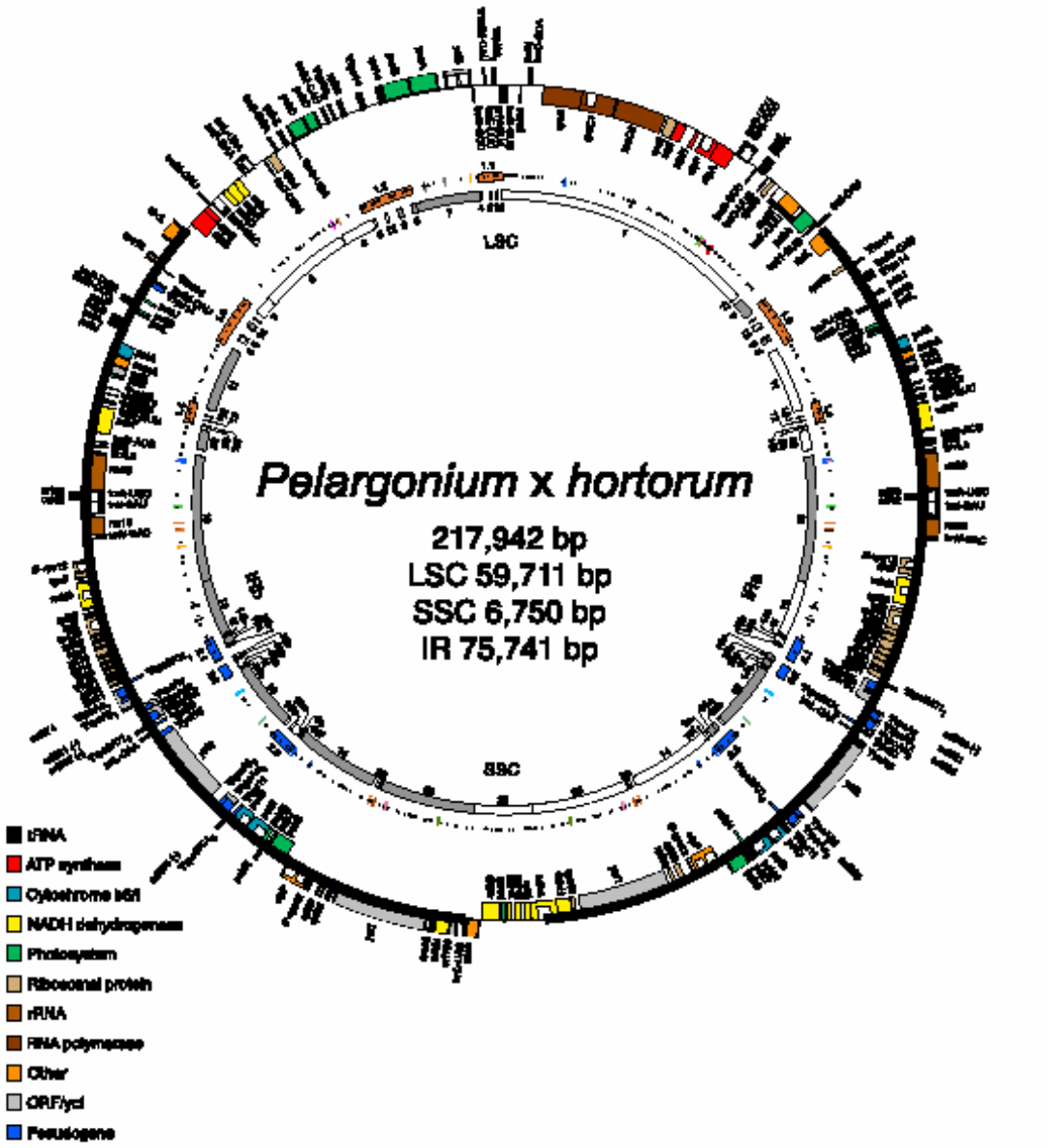
Fig. 4. Repeat association 1. a) Percentage identity plots from MultiPipMaker showing identities within and between each of the four repeat segments. b) Cartoon of alignments of each segment against itself and the others. c) Simplified schematic showing major repeat elements within each segment (genes, pseudogenes, and repeats *a-h*) and composite repeats (repeats *r1-r5*).

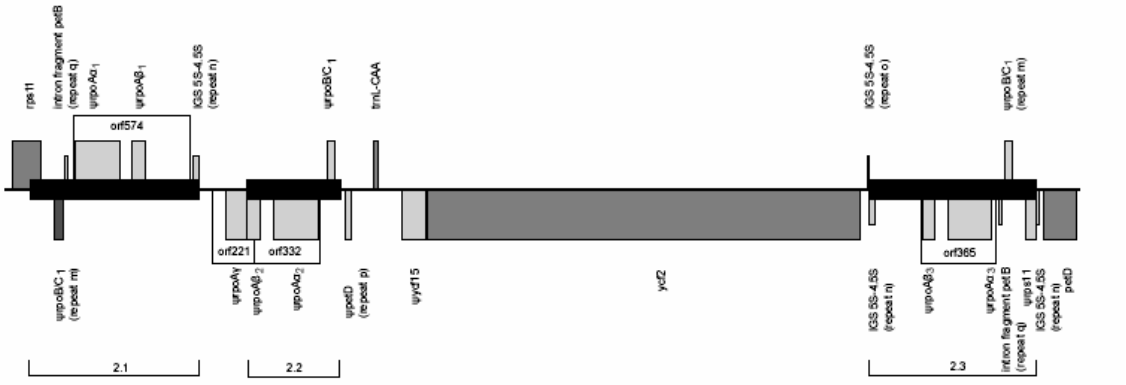
Fig 5. A simple evolutionary model for the major expansions and contractions of the IR and some of the inversions present in the chloroplast genome of *Pelargonium*. a) The presumed ancestral state. b) Small contractions of the IR remove *rpl2*, *rpl23*, and *ycf1* from the IR, leaving *trnI* at the IRa/LSC junction (JLA); inversions flip the order and orientation of *psbD-rps14* and *psaI-rps18*. c) A major contraction removes *trnL-trnI* (including *ycf2*) from the IR (leaving them only on the JLA side of the LSC) and an expansion into the SSC moves *ndhF* and *rpl32* into the IR; an inversion flips *psbD-psbZ* back into their original orientation, though appearing translocated, and another flips *rpl33-rps18* . d) Expansion of the IR into both the LSC and SSC including the *S10* operon (*rpl23-rpoA*, possibly to *petD*) and *ycf1-ndhA*, respectively. e) Expansion of the IR to include *ycf2*, leaving *trnI* stranded at the beginning of the IR. f) Large expansion of the IR to include *rbcL*; inversion of *trnN-ndhF*. g) 50 kb inversion of most of the IR. h) A resulting structural intermediate. i) Current structure of the genome showing locations of the high complexity major repeat associations I and II. Please note that this model does not account for rearrangements found in these regions (see fig. 6 and 7), nor does their appearance in the final stage here imply anything about the timing of their development.

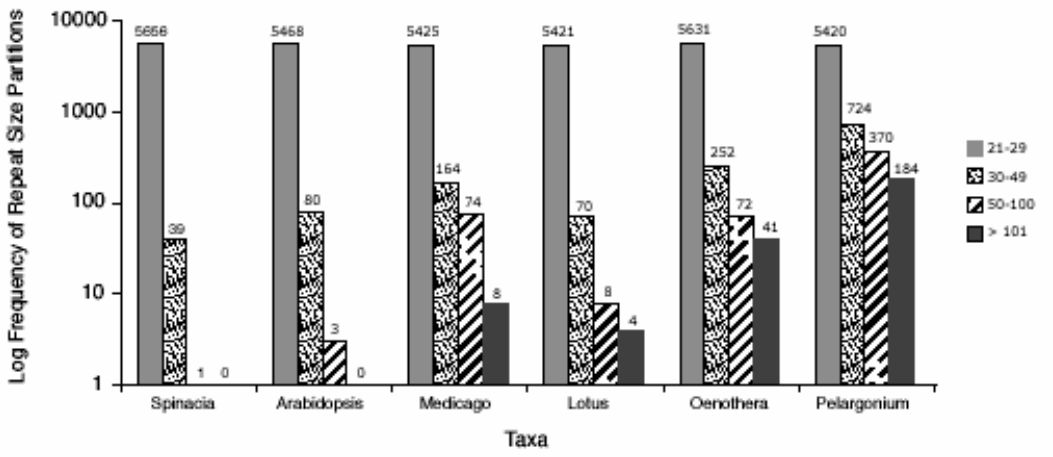
Fig. 6. An evolutionary model for major repeat association I. a) Putative ancestral arrangement of genes in this region, including duplications of *rpl33*, *trnfM* and *rps14*. b) A schematic diagram of the above, showing blocks of conserved gene order as found in the modern *Pelargonium* genome relative to tobacco. c-i) Inversion series required to transform putative ancestral genome into the modern. j) Schematic for the current

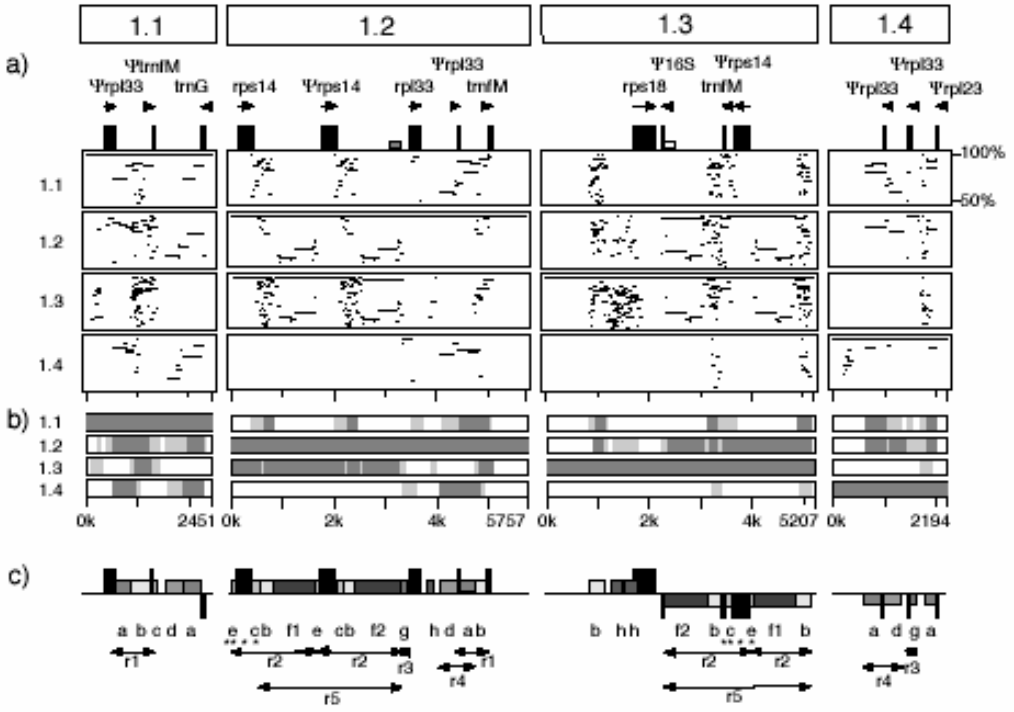
Pelargonium chloroplast genome. k) The current arrangement of genes for this region as determined in this study

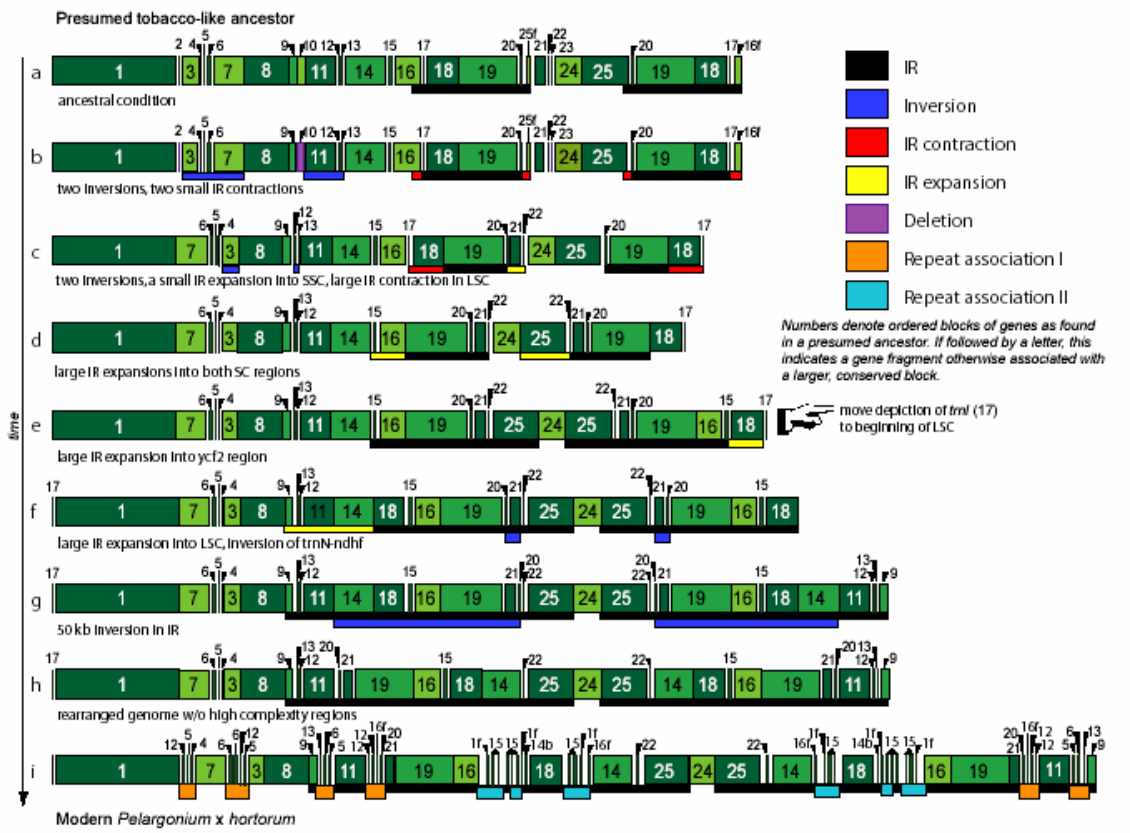
Fig. 7. An evolutionary model for major repeat association II.

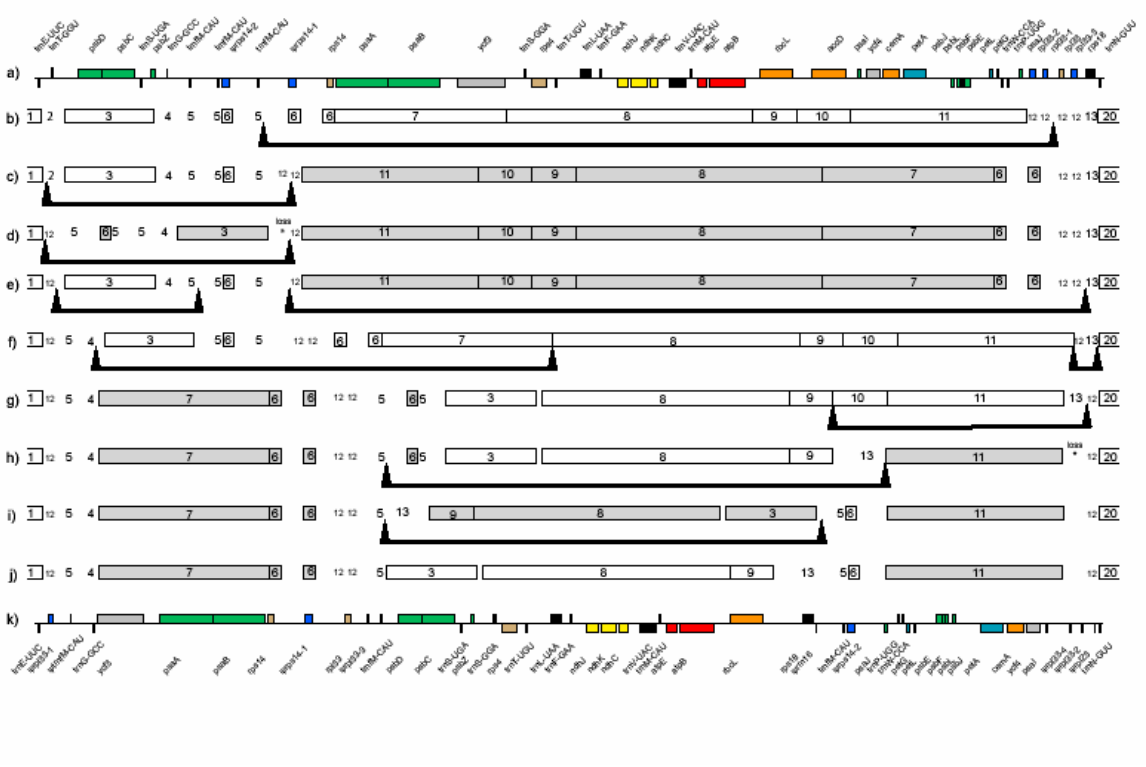

LSC
Pelargonium x hortorum
217,942 bp
LSC 59,711 bp
SSC 6,750 bp
IR 75,741 bp
SSC
tRNA
ATP synthase
Cytochrome b6f
NADH dehydrogenase
Photosystem
Ribosomal protein
rRNA
RNA polymerase
Other
ORF/ycf
Pseudogene

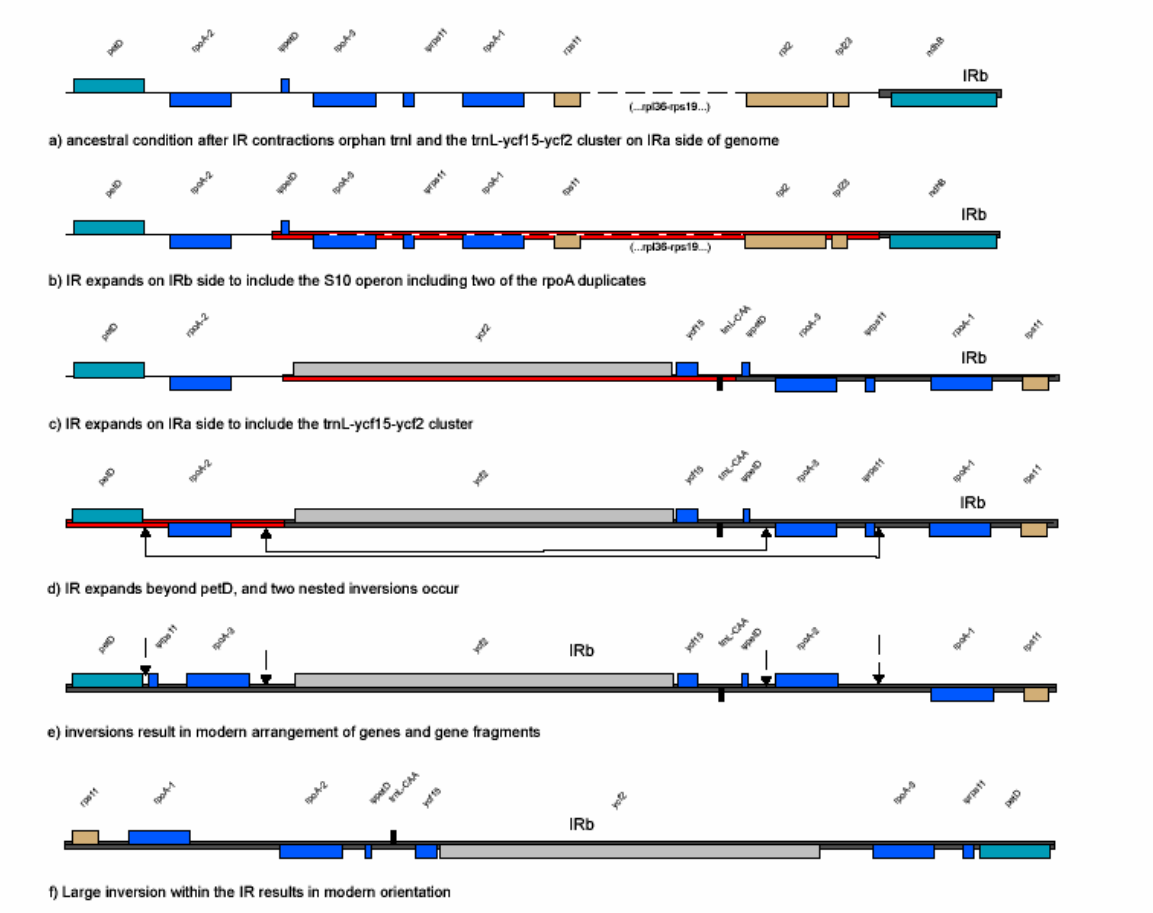

rps11
intron fragment petB (repeat q)
ψrpoA$_1$
ψrpoA$_1$
IGS 5S-4.5S (repeat n)
orf574
ψrpoB/C$_1$
trnL-CAA
IGS 5S-4.5S (repeat o)
ψrpoB/C$_1$ (repeat m)
orf221
orf332
orf365
ψrpoB/C$_1$ (repeat m)
ψrpoA$_2$
ψrpoA$_2$
ψpetD (repeat p)
ψycf15
ycf2
IGS 5S-4.5S (repeat n)
ψrpoA$_3$
ψrpoA$_3$
intron fragment petB (repeat q)
ψrps11
IGS 5S-4.5S (repeat n)
petD
2.1
2.2
2.3


Log Frequency of Repeat Size Partitions
10000
1000
100
10
1
5656
39
1
0
5468
80
3
0
5425
164
74
8
5421
70
8
4
5631
252
72
41
5420
724
370
184
Spinacia
Arabidopsis
Medicago
Lotus
Oenothera
Pelargonium
Taxa
21-29
30-49
50-100
> 101


1.1
1.2
1.3
1.4
a)
ΨtrnfM
Ψrpl33
trnG
rps14
Ψrps14
Ψrpl33
rpl33
trnfM
Ψ16S
Ψrps14
rps18
trnfM
Ψrpl33
Ψrpl33
Ψrpl23
100%
50%
1.1
1.2
1.3
1.4
b)
1.1
1.2
1.3
1.4
0k
2451
0k
2k
4k
5757
0k
2k
4k
5207
0k
2194
c)
a b c d a
r1
e cb f1 e cb f2 g
r2
r2
r3
h d ab
r4
f1
r5
b hh
f2 b c e f1 b
r2
r2
r5
a d g a
r4
r3


Presumed tobacco-like ancestor
a
ancestral condition
b
two inversions, two small IR contractions
c
two inversions,a small IR expansion into SSC, large IR contraction in LSC
d
large IR expansions into both SC regions
e
large IR expansion into ycf2 region
f
large IR expansion into LSC, inversion of trnN-ndhf
g
50 kb inversion in IR
h
rearranged genome w/o high complexity regions
i
Modern Pelargonium x hortorum
time
IR
Inversion
IR contraction
IR expansion
Deletion
Repeat association I
Repeat association II
Numbers denote ordered blocks of genes as found in a presumed ancestor. If followed by a letter, this indicates a gene fragment otherwise associated with a larger, conserved block.
move depiction of trnI (17) to beginning of LSC


a)
b)
c)
d)
e)
f)
g)
h)
i)
j)
k)


petD
rpoA-2
trnH-GUG
rpoA-3
trnT-UGU
rpoA-1
rps11
rpl2
rpl23
ndhB
IRb
(...rpl36-rps19...)
a) ancestral condition after IR contractions orphan trnI and the trnL-ycf15-ycf2 cluster on IRa side of genome
b) IR expands on IRb side to include the S10 operon including two of the rpoA duplicates
ycf2
ycf15
trnL-CAA
c) IR expands on IRa side to include the trnL-ycf15-ycf2 cluster
d) IR expands beyond petD, and two nested inversions occur
e) inversions result in modern arrangement of genes and gene fragments
f) Large inversion within the IR results in modern orientation